**Accelerated Strategic Computing Initiative (ASCI)**

# PathForward Project Description

**December 27, 1996**

**LAWRENCE LIVERMORE NATIONAL LABORATORY**

**LOS ALAMOS NATIONAL LABORATORY**

**SANDIA NATIONAL LABORATORY**

# ASCI PathForward Project Description

## Table of Contents

## I. Introduction

The Lawrence Livermore National Laboratory and the Los Alamos National Laboratory, operated by the University of California for the United States Department of Energy, and Sandia National Laboratories, operated by Lockheed-Martin for the United States Department of Energy, as part of the Department of Energy (DOE)'s Accelerated Strategic Computing Initiative (ASCI) are jointly initiating the PathForward Project, a set of multi-year development and engineering efforts to create the critical integrating and scaling technologies required for balanced computing environments at the scale of 10 to 30 TeraFLOPS (TFLOPS) in the late 1999 to 2001 timeframe.

The US commitment to ending underground nuclear testing, constraints on non-nuclear testing, and loss of production capability call for new means of verifying the safety, reliability, and performance of the US nuclear stockpile. One of these means is compute-based modeling, simulation, and virtual prototyping of nuclear weapon systems. The Accelerated Strategic Computing Initiative is one element of DOE's Stockpile Stewardship and Management Program and is designed to advance DOE's computational capabilities to help meet the future needs of stockpile stewardship. ASCI will create the leading-edge computational modeling capabilities that are essential for maintaining the safety, reliability, and performance of the stockpile. ASCI applications require near-term performance in the 10-to-30 TFLOPS range in the late 1999 to 2001 timeframe, and future developments enabling 100 TFLOPS platforms in the 2004 timeframe (see Figure 1). Additional background information on the ASCI program is provided at the end of this document and on the internet at http://www.llnl.gov/asci-pathforward.

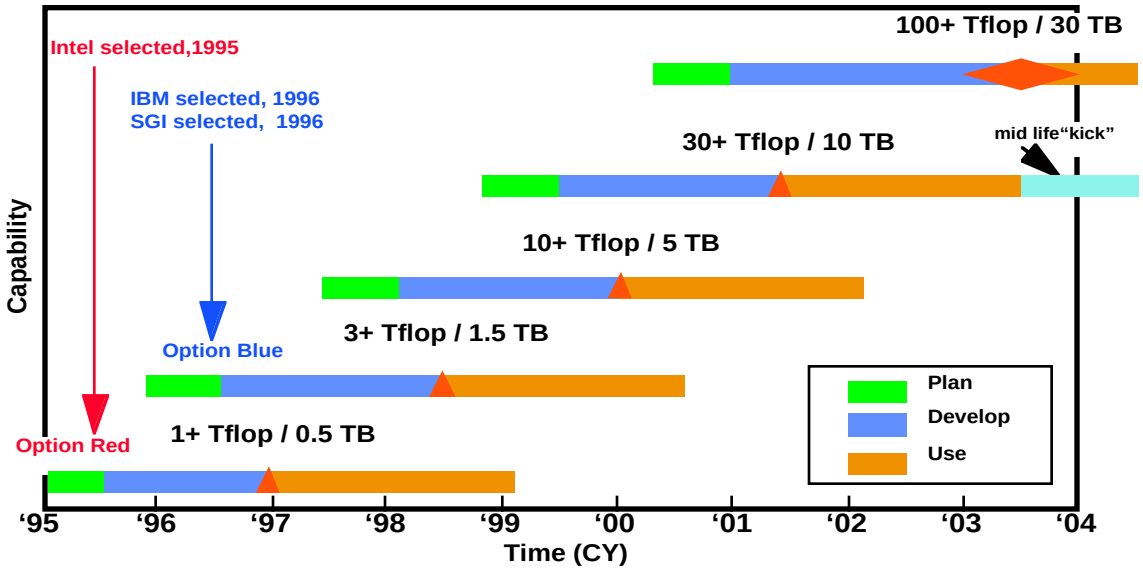



**Figure 1**: The ASCI roadmap for computing platforms.

Current trends in developing and implementing ultra-scale computers fall well below ASCI and Science Based Stockpile Stewardship Program mission requirements. ASCI applications require a threshold shift of 100 to 1000 times increase in computing capability. As a result, the U.S. Department of Energy and its National Labs are pursuing this PathForward Project to narrow the gap. Based on a number of factors, we believe that ultra-scale platforms will necessarily be developed using commodity -based components. This PathForward project is aimed at developing the leveraging components, technologies, and enhancements to commodity building blocks to attain the 100X to 1000X threshold shift that is vitally needed to meet mission requirements. Although this is the first phase, the PathForward project is an enduring concept that will be an integral component of the ASCI program through 2010.

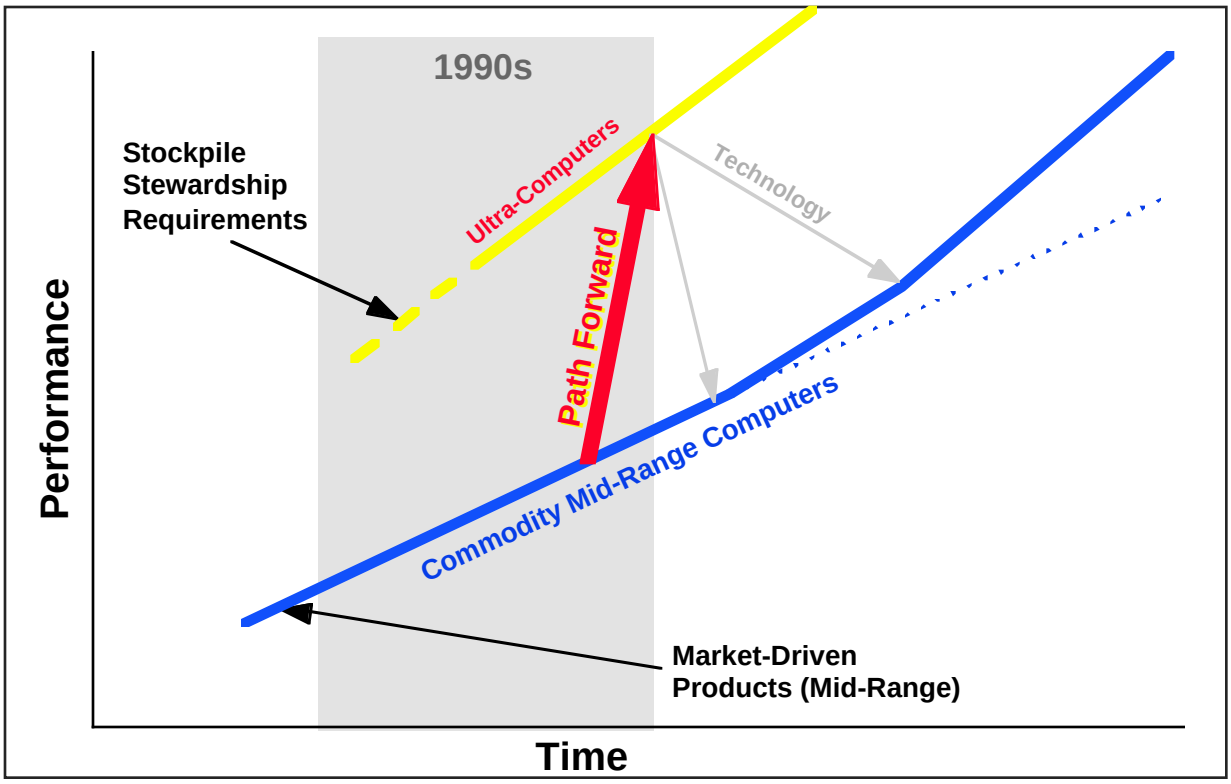



**Figure 2**: The PathForward Project

The aggregation of commodity building blocks into balanced 10 to 100 TFLOPS system environments will require significant development of integration and scaling technologies which are not currently being driven by commercial markets. The PathForward Project is intended to address an important subset of potential technology gaps for ASCI scale systems. This project description provides information about the initial scope of PathForward activities. At this time, we are particularly interested in initiating efforts which are relevant to the first phase of the PathForward Project, which targets 10-to-30 TFLOPS systems in the late 1999 to 2001 timeframe. Subsequent PathForward phases will address larger system scales and further timeframes, particularly 100 TFLOPS systems in the 2004 timeframe.

## II. 2004 ASCI Computing Requirements

The purpose of the PathForward Project's development and engineering efforts is to enhance the availability of the essential integrating and scaling technologies required to create a well-balanced, reliable and production capable computing environment providing large scale, scientific compute capability from commodity building blocks, at processing levels of 10-to-30 TFLOPS in late 1999 to 2001 timeframe and 100 TFLOPS in 2004. By a balanced environment we mean the total environment including scaleable hardware for processing, system level interconnections, networking, storage, desktop visualization and all the many levels of scaleable software necessary to provide the infrastructure for cost effective ASCI applications development and production use. This capability is needed to perform large scale numerical simulations in support of stockpile stewardship. At this time, we are particularly interested in industrial efforts with U.S. companies which will impact 10-to-30 TFLOPS systems. Subsequent PathForward Phases will address 100 TFLOPS systems. Further out in time, an even higher computational capability beyond the PetaFLOPS range is needed.

The simulation objectives, as described in the ASCI Program Plan, include nuclear weapon system performance, safety, reliability, and renewal. Simulations will be much more challenging because they will have to treat full 3D systems at high spatial resolution using predictive models based on fundamental physics. Simulations of aged or remanufactured weapon components will have to include effects due to microscopic material properties not previously understood or specified.

For the 2004 time frame, memory and speed estimates for these simulations are consistently in the range of 100 TBytes memory and 100 TFLOPS execution speed. To meet stockpile stewardship goals, 3D simulations and first principle physics modeling is required. For example, the radiation transport alone requires an estimated minimum of 40 TBytes memory and 123 TFLOPS speed for a representative simulation with a billion computational cells, assuming good scaling. For safety, 75 TBytes and 150 TFLOPS are needed for 3D simulations of shock waves propagating through explosive components, such as in a hypothetical complex transportation accident.

Aging simulations must address the effects of chemical changes, such as polymer breakdown, metal corrosion, and debonding, as well as effects from the decay of radioactive materials. Minimum estimates for simulations using methods such as molecular dynamics are 5.5 TBytes and 30 TFLOPS for studying highly localized phenomena, and grow rapidly as the scale of the simulation is increased. Detailed engineering simulations of microscopic degradation, fracture, and failure, have similar requirements in the 100 TBytes/100 TFLOPS range. Simulations of manufacturing processes such as casting, in which the microstructural material properties of the cast part must be predicted, have comparable computing memory and speed requirements.

## III. Achieving 100 TFLOPS ASCI Computational Performance

The technologies to be developed under the PathForward Project are based on an underlying approach in which the requisite computational resources are obtained by the aggregation of multiple commodity compute nodes. The purpose of the PathForward Project is the accelerated development of the unavailable, enabling non-mass market hardware and software technologies that will permit the creation of a balanced 100 TFLOPS compute environment by 2004 from mostly commodity hardware/software components. We recognize that the size and detailed implementation of commodity nodes, peripherals, and software is determined by market forces, and it is not a goal of the PathForward Project to drive the market sweet spot. This approach maximally leverages the naturally occurring technical progress driven by mass market forces. The PathForward Project is dedicated exclusively to the development of those hardware and software components which permit the scaling in computational capability by orders of magnitude beyond market-driven commercial systems by 2004. This concept includes the following key attributes:

- multiple high-performance commodity priced compute nodes, which represent regular commercial product lines and not special purpose designs;
- hierarchical memory systems, including cache-only memory architectures and distributed shared memory systems with low-latency high-performance memory access;
- commodity components and peripherals, such as DRAMs, disk and tape drives and controllers;
- supported by specially developed, essential integrating and scaling technologies, such as large-scale high-speed low-latency communication fabrics, very high performance storage and parallel I/O systems, scaleable programming environments and operating systems, a universal programming paradigm, etc.

The PathForward Project will concentrate its limited resources in those areas promising the biggest impact to removing critical path bottlenecks to ASCI-scale computing. For example, although we are very aware of the increasing processor-memory-bandwidth-latency-bottleneck and the large resources required to adequately address the problem, we are probably not in a position to impact this issue directly and will be relying on the mass market forces to address the problem. Also, this

problem does not seem to be specific just for the very high-end of computing, but applies to the commercial market in general.

Achieving balanced systems at the 10 to 100 TFLOPS scale is anticipated to place stringent requirements on the internode interconnect, I/O, and storage subsystems. Balanced systems suitable for ASCI applications are anticipated to scale according to the following approximate ratios:

1 TeraFLOPS peak performance /
1 Terabyte memory size /
50 Terabyte disk storage /
16 Terabyte per second cache bandwidth /
3 Terabyte per second memory bandwidth /
0.1 Terabyte per second I/O bandwidth /
10 Gigabyte per second disk bandwidth /
1 Gigabyte per second archival storage bandwidth /
10 Petabyte archival storage.

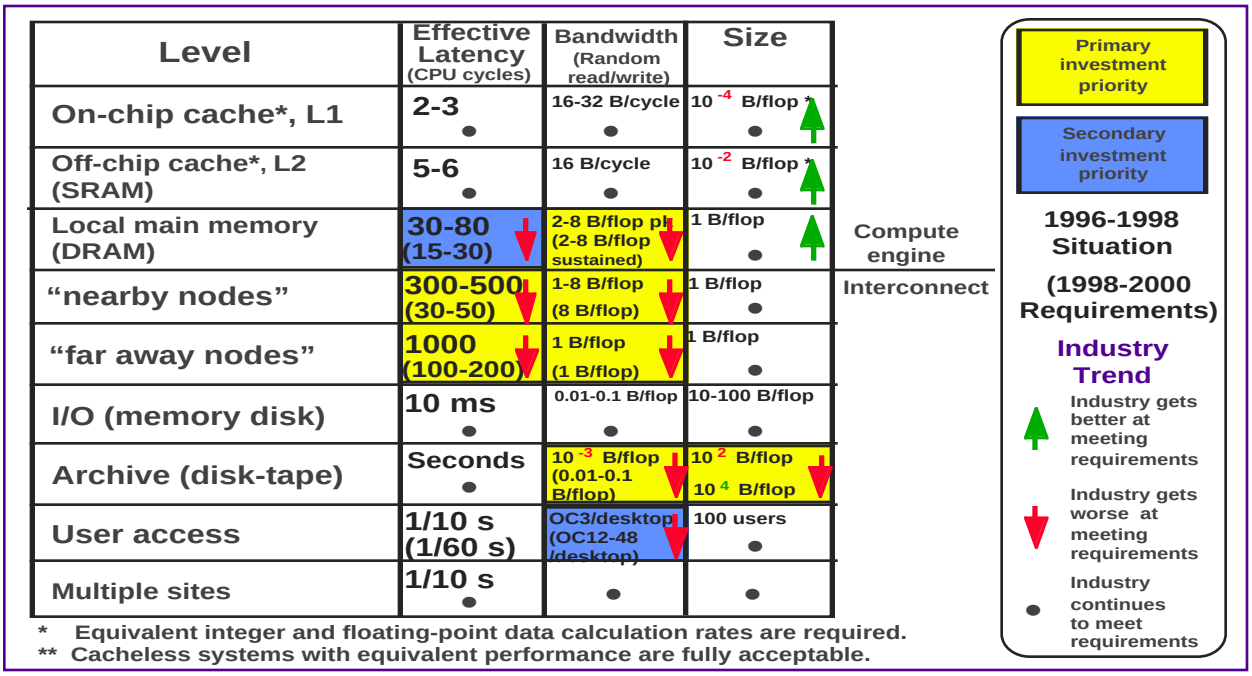
We recognize that, in scaling up over several orders of magnitude, these goals will be difficult to achieve in ultra-scale systems, and the archival storage cost may be prohibitive.

To facilitate the applications development efforts for stockpile stewardship and to enable ease of use in a multitasked programming environment, the first phase target 10-to-30 TFLOPS systems will require significantly enhanced software. Software in both the distributed operating system and programming environments is critical to our ability to utilize 100 TFLOPS platforms effectively for stockpile stewardship. Project effectiveness is expected to occur through

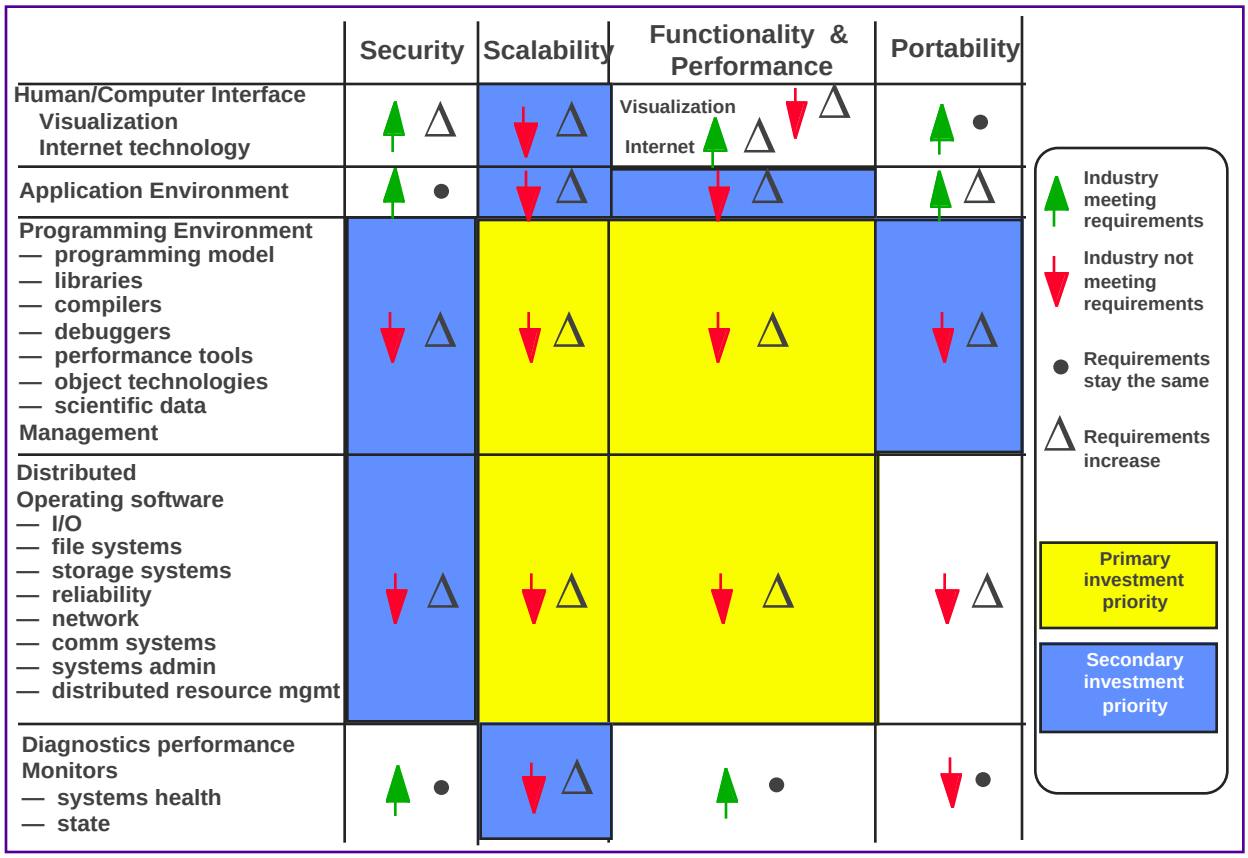
- system integration;
- tools enabling efficient programming and load balancing; and
- other elements required to present a single-system view of the hardware resources and to enable the efficient use of these resources.

Our current perceptions of commercial technology trends, and how they relate to anticipated ASCI performance requirements, are summarized in Tables I (hardware) and II (software). In Table I, hardware requirements are summarized in the context of an hierarchical memory access model, using normalized measures (CPU cycles, bytes/FLOP) where appropriate. In both tables, a preliminary identification of technology areas targeted as primary investment priorities is indicated by the shaded regions. Further detail on these areas is provided in the following “PathForward Project: Development and Engineering Opportunities" section.

## Table I: ASCI Hardware Requirements and Technology Trends




## Table II: ASCI Software Requirements and Technology Trends



## IV. PathForward Project: Development and Engineering Opportunities

The PathForward Project's goal is to accelerate the development and availability of those integrating and scaling technologies which enable companies to leverage commodity computing building blocks into usable systems for scientific computing at 100 TFLOPS and beyond. This first phase of the project specifically targets technologies that can impact 10-to-30 TFLOPS systems in the late 1999 to 2001 timeframe. The PathForward Project will accelerate only those essential integration and scaling technologies, rather than development of the commodity building blocks themselves. Modification of commodity building blocks to enable their use in ultra-scale systems is also of interest. In particular, this multi-year project will invest in those hardware and software technologies that leverage the success of the commercial microprocessor and peripherals industry:

- to achieve balanced, high-end, scientific computing at 10 TFLOPS and beyond,
- that can be made available on a time scale sufficiently rapid to impact future ASCI platform and subsystem procurements--notably 30 TFLOPS in the 2001 timeframe,
- that will accelerate extended capabilities of commercial product lines to market, and
- will not be created by commercial market forces in the time scale of interest.

The PathForward Project must significantly accelerate the development of key elements of the balanced hardware/software computing infrastructure required for 10-to-30 TFLOPS scale computing ahead of planned product offerings. These capabilities should be on the commercialization path of the selected participants, not one time offerings.

The PathForward Project will also be exploring efforts to develop non-proprietary (pre-competitive) technologies and standards to facilitate integrating systems in areas considered non-competitive by the industry. While there are many areas of computing infrastructure which must be advanced to meet our goals, initial funding will be available only for elements which are fundamental to achieving balanced, 10-to-30 TFLOPS computing environments. The following paragraphs describe elements which we have identified to be of primary concern. Initial efforts will not necessarily be limited to these areas. We welcome alternatives of equivalent import.

### Scaleable High-Performance Interconnects

The 10-to-30 TFLOPS computing system necessary for stockpile stewardship in the late 1999 to 2001 timeframe will require the interconnection of order $10^4$ CPUs, or equivalently of order $10^3$ computing nodes. The performance of the interconnection between computing nodes, relative to that within such a node, directly impacts the ease of programming such systems to attain a high level of sustained performance on the applications of interest. Achieving high performance becomes significantly more difficult with increases in remote (inter-computing node) memory access time. For this reason, the ideal system interconnect should provide a bisection bandwidth of at least 5 to 15 TBytes/s and a latency of a few hundred CPU cycles (including software overhead). It is unclear, however, that straightforward extensions of current interconnect technologies will prove satisfactory due to increased congestion and latency as well as the aggregate bandwidth not scaling with system size. Innovative approaches will be required for large scale interconnects, which build upon the infrastructure developed by vendors for moderate size computing clusters. Several sample areas for development opportunities are listed below. Other approaches leading to high performance, large scale system interconnects in the near to medium term are also of interest.

Opportunities exist to develop and demonstrate prototypes of an interconnection fabric with sufficient bandwidth, latency, and scalability to support the balanced 10-to-30 TFLOPS systems described in this document. This includes the technologies required for low overhead node-to-interconnect access, high performance data transport, and scaleable interconnect architecture.

We are also interested in developing open standards for high performance node-to-interconnect interfaces and interconnect components, and to have prototype implementations and demonstrations of these standards. Standardized interfaces with low overhead and latency will enable the use of compute nodes from different companies with a common interconnect, or vice versa, allowing the separate optimization of the nodes and interconnect selected for ASCI computing platforms.

We are also interested in the development and demonstration of interconnect prototypes which support advanced features, implemented in hardware or low-level software, which alleviate the programming burden required to achieve efficient performance on 10-to-30 TFLOPS scale systems. Examples of such features are:

- extension of cache-coherent, shared memory technology to larger numbers of compute nodes, so that these nodes present a single system image
- cache-only memory architectures, to enable the local storage of larger data sets than can be accommodated in conventional node caches.

Distributed Parallel Operating System Software

We envision an ultra-scale computing environment in which a number of computers or nodes are integrated through the aforementioned, high-performance interconnect fabric into a single effective multi-user, high-end computing system, supporting both capability and capacity computing of 10-to-30 TFLOPS in the late 1999 to 2001 timeframe. Such a system should present itself to the user through a standardized single system image and be operational with integrated system-wide resource allocation and management. This places tremendous challenges on a distributed (shared memory) parallel operating system from both a scalability and portability point of view.

We are interested in efforts to adapt and develop a scaleable, portable, standardized, distributed parallel operating system architecture and software which sits on top of the "shrink-wrapped" operating systems of the individual computers or nodes. We envision an approach similarly to the one taken by the Data Management Interface Group (DMIG) with standard OS API extensions (standard hooks) supported by multiple sources.

In our case these standard OS API extensions will be a requirement for the "shrink-wrapped" part of the operating system of any computer or node to be integrated into our high-end computing environment. The portable standardized distributed parallel operating system should be scaleable to 100s to 1000s of nodes with a total number of 10,000s of CPUs.

Particular areas of interest to be built on the OS extension API are: data movement protocols which can adapt to and take advantage of interconnection network advances; fault tolerance mechanisms which improve system reliability and allow applications to survive faults; accelerators of standard application-level APIs such as MPI and POSIX threads; mechanisms to improve global resource management through the ability to submit jobs and control allocations of all resources CPUs to memory to disks to archival storage; and mechanisms to improve security of data and of user access.

Parallel Scaleable Programming Environment: Debuggers and Performance Monitors

In a balanced computational environment, hardware and software improvements have contributed equally to the increase in applications performance. What matters most is the shortest time to solution. Code development efforts are definitely on the critical path to success. We are focusing on reaching ultra-scale performance levels, suitable for solving large-scale complex problems. The fully integrated heterogeneous, multi-vendor parallel computing environments we are considering pose considerable challenges from an application

programming point of view. Code development efforts addressing the hierarchical distributed shared memory programming paradigm, which is a functional representation of this computing environment, have to rely in an essential manner on easy to use, high-performance debugger and performance monitor tools supporting the collection of computer systems in its entirety. This support has to include message passing and shared memory environments, as well as a combination thereof.

Opportunities exist to adapt and develop an integrated, scaleable, high-performance parallel debugger and performance monitor architecture and software which are portable and function effectively across distributed platforms with up to approximately 10,000 processors, corresponding to millions of simultaneous processes or threads of control. These tools have to be able to relate debugging information back to the source code. Because these tools are intended to be used on 10-to-30 TFLOPS systems with up to 10-to-30 TBytes of RAM, a similarly large amount of data must possibly be dealt with in a debugging context. Ease of use, graphical representations of data in multiple dimensions, and high-performance parallel data access by these tools seem to be essential ingredients. The debugger and performance monitor tools have also to support MPI and POSIX threads. The software has to function on top of the distributed parallel operating system software described above. However, this activity is not intended to solve distributed computing or the problem solving environment.

Ultra-scale, High-Performance Storage

We are potentially interested in working with industry on storage technology, subsystems and total system issues, provided that suitable approaches can significantly advance our ultra-scale goals. It is not sufficient to just develop faster parallel processing systems and local memories. If ultra-scale systems are going to achieve their potential to enable applications which are not currently possible, then these systems must be embedded in a total balanced environment including local disk storage, high speed networking, tertiary storage and visualization at the desktop. As processing scales to 100 TF and beyond tertiary storage capacities must scale to Exabytes with bandwidth to the processing and local disk systems in the 100s of GB/s region and beyond. The tertiary storage system is likely in the timeframe of this project to consist of both disk for caching and tape for long term storage. The problem is how to scale tertiary storage systems using commodity disk, tape, robotics and other emerging tertiary storage media and systems to meet the above capacity and performance goals reliably, in an acceptable footprint and at an affordable cost. Example problem areas for scaling existing technologies are listed below. However other approaches that could be made available in meeting the balanced needs of HEC storage systems in the near to medium term are also of interest.

Exploration of alternative architectures for reliable, high density, high performance parallel tape subsystems (RAITs), controllers and robotics, and the demonstration of one or more prototypes is of interest. In order to implement tertiary storage systems that scale in size to Exabytes (with reasonable footprint and cost) and in I/O performance to 10s - 100s of Gigabytes/sec and beyond will require improvements in density and single tape I/O performance (e.g. optical tape might be a useful direction). These individual units will have to then be integrated into parallel tape subsystems with some form of error detection and correction mechanism such as parity across striped tapes. There will also have to be associated robotics libraries that can in a single robotics movement store and retrieve the multiple cartridges of a parallel tape volume and perform loads of parallel read/write stations in a single operation.

Similarly, opportunities exist to explore much wider RAID architectures and build and demonstrate one or more prototypes that will scale as individual units into the 1-10 Gigabyte/sec range. In order to scale using these RAIDs or the above RAITs into higher order parallel configurations, opportunities also exist to develop and demonstrate prototype, high

speed, interconnection fabric- and network-attached-peripheral (NAP) controllers for peripheral systems. NAPs would support third-party control of data transfer and also authenticate and authorize access; as well as contribute to industry standards work on NAPs. Implementation of scaleable, cost effective, high speed, large scale storage systems of 10s - 1000s Petabytes will require NAP based architectures.
(see www.llnl.gov/liv_comp/siof/hpss_nap_wg.html)

In summary, we have identified four development and engineering opportunities of particular interest: scaleable interconnects, distributed parallel operating system software, tools for a parallel scaleable programming environment, and ultra-scale storage. Development activities leading to high-performance, scaleable interconnect fabrics are currently our highest priority. We are potentially interested in working with industry on all of these issues, provided that suitable approaches can significantly advance our ultra-scale goals.

## V. Implementation

We anticipate establishing more than one distinct effort with U.S. companies providing high performance computing technology in the first phase of the PathForward Project. Different efforts may involve different companies, and may focus on different technology areas. At this time, the funding anticipated for these initial efforts is approximately $10 million per year, with anticipated project durations of two to four years. We envision the PathForward first phase efforts to emphasize the development of hardware and software systems that can achieve performance at the 10 to 30 TFLOPS level, and middleware components necessary to develop and execute the applications of interest for stockpile stewardship. We are particularly interested in efforts for which a viable path exists to incorporate the technology development into either complete high end computer systems or in building blocks that may be used in building high end systems.

Teaming arrangements among a group of suppliers are encouraged, for example consortia that address innovative integrated software and/or hardware approaches to providing high end computing solutions composed of commodity market building blocks. Teaming arrangements that provide open systems technology and industry standards are also encouraged.

While the primary PathForward thrust will be accelerated technology development, we require the delivery of tangible results from the proposed work. These results should be embodied in a sequence of incremental deliverables, determined by the computer companies and National Labs, with verifiable metrics distributed over the multi-year performance period. In addition, researchers from the participating National Laboratories should have ready access to prototypes or useable, pre-commercial components for evaluation and verification purposes.

We anticipate that intellectual property rights to inventions developed under PathForward funding may be retained by the company(ies) performing the development, rather than by the government, provided that appropriate conditions exist and documentation is filed with the U.S. Department of Energy.

The spirit of ASCI is “One Program--Three Labs”. All three Defense Programs Laboratories (Lawrence Livermore National Laboratory, Los Alamos National Laboratory, and Sandia National Laboratories) will participate in the technical oversight of all PathForward activities and the application of technical resources developed under the PathForward Project.

# Appendix: ASCI Background Information

A basic element of the ASCI high-end computing strategy is the creation of multiple industrial research and development projects to accelerate current trends in high-performance computing. It is not possible to predict with confidence which of several competing technology paths will lead to the best results; indeed, it is unlikely that a single technical approach will serve all program needs. Therefore, the R&D efforts overlap both in time and technical approach in order to achieve high payoff with manageable risk, to ensure flexibility, to avoid dependence on single sources, and to serve the wide diversity of ASCI application environments.

Platforms to meet the near-term ASCI applications requirements are being obtained via a set of contracts with individual computer companies. These projects, known as "Option Red" and "Option Blue", are providing computational platforms at each of the national laboratories ranging from one to three TFLOPS peak performance. A key feature of these platforms is their aggregation of "commodity off-the-shelf" technology in scaleable architectures to attain high-performance systems at affordable costs. This strategy enables the National Laboratories to take advantage of the R&D investments that industry makes in its commercial product lines and apply the results to the very high performance computing requirements of the Stockpile Stewardship Program. It leverages the price/performance advantage of both hardware and software commodity items, and also enables the leveraging of a broad applications development tool base established by independent software vendors.

The aggregation of commodity building blocks into 10 to 100 TFLOPS systems requires additional, significant development of integration and scaling technologies which are not currently being driven by commercial markets. The PathForward Project is intended to address the potential technology gaps resulting from these circumstances. It targets the near-term development and acceleration of those specific, essential technologies which will be required to implement 100 TFLOPS platforms from commodity building blocks on the 2004 timescale. The focus of the acceleration strategy is therefore on integration and scaling technologies, rather than on the individual commodity building blocks.

Additional information on the ASCI program may be found on the internet ASCI home page, at http://www.llnl.gov/asci.